LBSN 协作式个性化链接预测算法 *

胡　敏[1,2]，崔永胜[1]，黄宏程[1†]，陈元会[1]

(1. 重庆邮电大学 通信与信息工程学院，重庆 400065; 2. 重庆市通信软件工程技术研究中心，重庆 400065)

摘　要：在基于位置的社交网络中用户链接与位置链接之间具有一定的内在关联，而且不同的用户在社交网络中表现也存在差异。因此，对于以上问题提出一种协作式个性化链接预测算法。针对用户的个性化特征，采用核密度估计方式对用户在时间和空间维度建模，基于兴趣组对用户进行重叠社团划分，并通过社团、好友以及签到关系进行个性化用户链接预测。基于个性化用户链接预测结果，利用从社团重启的随机游走预测用户的个性化位置链接。协作式个性化链接预测算法通过用户链接预测和位置链接预测的迭代使得二者性能相互提升，实验结果表明，所提算法相比于现有算法具有更好的预测性能。

关键词：链接预测；基于位置的社交网络；核密度估计；个性化；随机游走

中图分类号：TN915.07　　doi: 10.19734/j.issn.1001-3695.2018.10.0768

Cooperation based personalized link prediction algorithm in LBSN

Hu Min[1, 2], Cui Yongsheng[1], Huang Hongcheng[1], Chen Yuanhui[1]

(1. *School of Communication & Information Engineering*, *Chongqing University of Posts & Telecommunications*, *Chongqing* 400065, *China*; 2. *Chongqing Engineering Research Center of Communication Software*, *Chongqing* 400065, *China*)

Abstract: There is a certain internal relationship between user links and location links in location-based social network(LBSN) , and different users also have different behaviors in the network. Therefore, view of the above problem, a Cooperation based personalized link prediction algorithm(CPP) is proposed in LBSN. For the user's personalized features, the kernel density estimation method is used to model the user's time and spatial dimensions. The interest groups were used to divide the users into overlapping communities, and the personalized user link prediction was performed through the community, friends and sign-in relationships. Based on the prediction of the personalized user link, a personalized link relationship between users and locations was predicted via the algorithm of the random walk with community restarting. The CPP algorithm improves the performance by the iteration of the user link prediction and the location link prediction. The experimental results show that the CPP algorithm has better prediction performance than that of the existing algorithm.

Key words: link prediction; location-based social network; kernel density estimation; personalization; random walk

0　引言

随着社交网络的快速发展和移动智能终端的不断普及，基于位置的社交网络(location-based social network, LBSN)逐渐成为人们维系社交关系、分享位置信息的理想网络平台[1]。越来越多的用户习惯使用智能终端在社交网络平台上进行位置签到，但是随着用户数的增加令网络数据量爆发增长，从而导致网络信息过载[2]。LBSN 中的链路预测研究可以帮助用户从海量数据中发现潜在的用户链接关系，并推荐用户感兴趣的其他用户或者位置信息，对于把握 LBSN 结构的演化规律，增加用户对 LBSN 平台的忠诚度等方面具有重要的研究意义和应用价值[3]。

链接预测根据网络结构以及网络中已有信息发现并且还原网络中缺失的信息，或者预测未来节点之间可能存在的关系，其研究对于好友推荐，兴趣点推荐等应用具有重要的现实意义[4]。

目前 LBSN 中的链接预测主要分为两类，一类是预测用户与用户之间的链接，另一类是预测用户与位置之间的链接。针对用户与用户链接预测问题，Valverde-Rebaza 等人[5]考虑用户之间的关系强度和用户位置信息，结合用户的社交模式和移动模式来提高链接预测的准确性。丁勇等人[6]提出从兴趣、距离和熟识度三个属性构建好友推荐模型，此外，还考虑了用户的交友偏好属性。Bayrak 等人[7]提出不同类别位置对于链接建立的影响程度不同，提出两种新的基于类别的特征，从而提高用户的链接预测性能。针对用户与位置链接预测问题，Pavlos 等人[8]考虑了用户评论的社会影响和用户签到的空间影响特征，通过考虑这两个特征，预测用户与位置的链接关系。李鑫等人[9]提出了一种在 LBSN 上基于兴趣圈社会关系模型，使用社会关系包含朋友关系和专家用户，通过这两个规则化项作为矩阵分解目标函数的约束项，来提高预测用户与位置的链接性能。Hosseini 等人[10]认为用户和位置之间应当存在一种对应关系，即如果用户喜欢在工作日活动，那么应当给用户推荐工作日受欢迎的位置，同理，对于喜爱周末活动的用户，应当给其推荐周末受欢迎的位置。

收稿日期：2018-10-13；**修回日期：**2018-12-19　**基金项目：**国家自然科学基金（61871062）；重庆邮电大学科研基金（A2018-07）

作者简介：胡敏（1971-），女，重庆人，副教授，硕士，主要研究方向为虚拟现实、脑机接口、通信网体系协议；崔永胜（1992-），男，河南临颍人，硕士研究生，主要研究方向为社会计算、大数据技术及应用；黄宏程（1979-），男（通信作者），河南南阳人，副教授，博士，主要研究方向为复杂网络与信息传播理论、社会感知与智能计算（huanghc@cqupt.edu.cn）；陈元会（1991-），男，湖北武人，硕士研究生，主要研究方向为社会网络信息传播与控制.

上述方法分别以相互独立的目的来预测用户与用户链接关系以及用户与位置的链接关系，没有考虑二者之间的相关性，然而现实中两者之间并非毫无关联。例如，如果两个用户之间存在链接关系，则他们很有可能在相同的位置签到。另外，如果两个用户经常在相同的位置签到，则他们很有可能存在链接关系。因此，用户链接和位置链接之间存在较强的关联性，并且能够相互促进彼此链接预测性能。目前，将这两个问题联合解决的工作还很少。Zhang 等人[11]提出一种新的链接预测方法 TRAIL，通过最大化用户链接预测和位置链接预测乘积，得出最优用户链接和位置链接关系。文献[12]提出一种锚链接预测方法，锚链接预测包括用户的社交，空间和文本信息预测。

在现实生活中，地理位置的邻近性对于用户的签到行为有着显著影响，上述研究虽然有些考虑到了用户与位置链接预测的协作性，却未能有效融合地理位置信息对预测性能的影响。目前有以下两种考虑位置信息影响的方法，第一种是根据用户与位置的距离远近，过滤到距离用户比较远的位置[13]；第二种是将用户的签到数据建模为一种概率分布函数[14]。第二种方法对于地理位置信息使用考虑更加严谨，所以得到的链接预测性能也较好。但是，不同的用户对于空间位置的容忍度不同，建立统一的概率分布模型掩盖了用户的个性化特征，使得用户个性化特征丢失，影响预测准确度。另外，用户的签到习惯也不相同，不同的用户喜欢出去的时间也不同，通过更加匹配用户的习惯将会进一步提高算法的预测性能。因此，本文通过对每一个用户进行个性化建模，更加准确的把握不同用户在空间位置和行为习惯上的个性化特征。

首先，考虑用户链接预测和位置链接预测的协作性问题，再者考虑不同用户的个性化特征。本文提出一种 LBSN 中基于协作式的个性化链接预测算法(cooperation based personalized link prediction，CPP)，从一种新的角度提高基于位置的社交网络链接预测性能。

1 问题描述

基于位置的社交网络可以视作是由不同类别节点和边组成的异构网络，本文使用三元组 $G=(V,E,A)$ 来表示，其中，V 表示节点集合，E 表示边集合，A 表示节点类型集合。通过几个相关定义来更好的说明问题。

定义 1 兴趣组。假如用户 u 和用户 v 都在类别为 c 的位置签到，则定义用户 u 和用户 v 属于同一个兴趣组 c。

定义 2 本地位置重要度(local location importance，LLI)。给定用户 $u\in \mathrm{U}$，U 表示所有用户集合。令 $p\in \mathrm{P}_u$，P_u 为用户 u 访问过的位置集合，N_u 表示用户 u 的签到总次数，$n_u(p)$ 表示用户 u 在位置 p 签到的次数。本地位置重要度为位置 p 相对于用户 u 访问过的所有位置的重要度，公式为

$$\mathrm{LLI}_u(p)=1/(-\log(\frac{n_u(p)}{N_u})) \tag{1}$$

定义 3 全局位置重要度(global location importance，GLI)。N_p 表示所有用户在位置 p 签到的总次数，其他符号定义与上述相同。全局位置重要度为用户 u 相比于其他用户对位置 p 的重要度，公式为

$$\mathrm{GLI}_u(p)=1/(-\log(\frac{n_u(p)}{N_p})) \tag{2}$$

为了形式化地描述本文研究的科学问题，本文将 LBSN 建模为带权异构网络形式，使用四元组 $\mathrm{G}_h=(\mathrm{U}_h,\mathrm{P}_h,\boldsymbol{E}_h,\mathrm{W}_h)$ 表示。其中：$\mathrm{U}_h=\{u_1,u_2,\ldots,u_n\}$ 表示用户节点集合。

$\mathrm{P}_h=\{p_1,p_2,\ldots,p_m\}$ 表示位置节点集合。

$\boldsymbol{E}_h=E_{uu}\cup E_{up}\cup E_{pu}$ 表示网络中的边集合，本文中所描述的边为有向边。

$\mathrm{W}_h=W_{uu}\cup W_{up}\cup W_{pu}$ 表示网络中边权值集合。

首先根据相关定义划分用户群体，构建初始预测空间 C。然后，通过本文提出的链接预测算法，预测网络中可能存在的用户链接 E_U^+ 和位置链接 E_P^+，相关问题定义表示为

$$\left.\begin{matrix}\mathrm{G}_h(\mathrm{U}_h,\mathrm{P}_h,\boldsymbol{E}_h,\mathrm{W}_h)\\ \mathrm{C}\end{matrix}\right\}\Rightarrow f:(\mathrm{U}_h,\mathrm{P}_h,\boldsymbol{E}_h,\mathrm{W}_h)\rightarrow\left\{\begin{matrix}\mathrm{E}_U^+\\ \mathrm{E}_P^+\end{matrix}\right. \tag{3}$$

1.1 问题输入

基于上述定义，本文研究内容的输入为：

a) 带权异构网络 $G=(U,P,\boldsymbol{E},W)$；

b) 初始预测空间 C。

1.2 问题输出

在给定基于位置的社交网络中的带权异构网络 $\mathrm{G}_h=(\mathrm{U}_h,\mathrm{P}_h,\boldsymbol{E}_h,\mathrm{W}_h)$ 以及初始预测空间 C 的前提下解决问题如下：

a) 如何对用户个性化建模？使用非参数估计方法中的核密度估计，针对每个用户签到时间以及用户空间位置容忍度个性化建模。优点是不需要提前假定样本的分布特性，适用于小样本数据集。时间建模使用针对时间标量的一维核密度估计方式，空间建模使用针对经纬度坐标向量的二维核密度估计方式。

b) 如何设计用户个性化链接预测算法，并同时解决基于位置社交网络中的用户链接预测和位置链接预测问题？根据用户历史签到位置得到用户兴趣组，再依据兴趣组对用户进行社团划分。利用用户链接预测方法计算两个用户间链接的概率，并使用用户时间维度的个性化特征更新用户链接概率。通过社团重启的随机游走得到每个社团中的重要位置，并通过用户空间维度的个性化特征更新每个用户的位置链接概率，最后通过迭代用户链接以及位置链接，从而使二者性能相互提升，得到本文的预测结果 $\mathrm{E}_U^+\cup\mathrm{E}_P^+$。

2 用户个性化建模

目前大量研究都是使用全局用户数据进行建模，极少考虑用户个性化行为习惯和个人喜好，从而造成用户个性化信息损失。本文根据用户的签到数据，分别从时间和空间两个角度对单个用户进行个性化建模，从而获取用户个性化特征，进一步提高用户链接预测和位置链接预测的准确性。本文采用核密度估计方法来对用户进行个性化建模。

2.1 核密度估计

核密度估计是一种非参数的估计方法，它的优点是不需要提前假定样本的分布，可以根据样本本身发现其分布特征。相比于参数估计方法，避免了复杂的分布假设以及参数回归过程，使得估计样本分布变得简单高效，因此，核密度估计方法很适用于对单个用户个性化建模。根据样本对象数据维度不同，可以将核密度估计分为一维核密度估计和多维核密度估计。

2.1.1 一维核密度估计

假设 $(x_1,x_2,\cdots,x_n)$ 是取自于一个独立同分布样本的随机变量集合，f 表示其未知的概率密度函数。则其核密度估计公式为

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{k}_h(x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}\mathrm{k}(\frac{x-x_i}{h}) \tag{4}$$

其中：$h>0$ 表示窗宽值，$\mathrm{k}(\cdot)$ 为核函数，且核函数需要满足以下条件：

$$\int \mathrm{k}(u)d_u=1,\quad \int u\mathrm{k}(u)d_u=0 \tag{5}$$

2.1.2 多维核密度估计

当样本对象从标量形式转换成 q 维向量时，样本的密度分布函数就变为多维核密度估计。假定有 n 个 q 维的随机变量 $(x_1,x_2,\cdots,x_n)$，且服从独立同分布概率密度函数 f 。则其多维核密度估计为

$$\hat{f}(x)=\frac{1}{nh_1h_2\ldots h_q}\sum_{i=1}^{n}\mathrm{K}(\frac{x-x_i}{h}) \tag{6}$$

其中:多维核密度估计的核函数由一维核函数的乘积构成

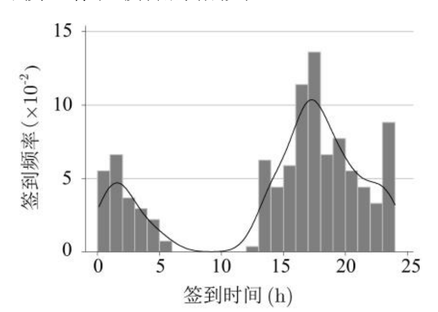
$$\mathrm{K}(\frac{x-x_i}{h})=\prod_{i=1}^{q}\mathrm{k}(\frac{x-x_i}{h_i}) \tag{7}$$

当 q=2 时，即为二维核密度估计函数。

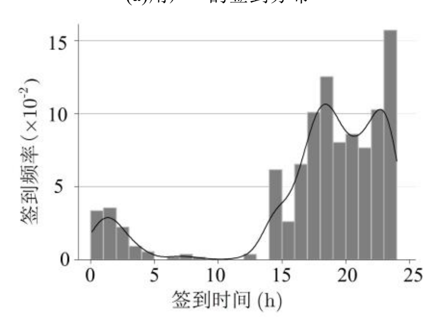
2.2 用户个性化时间建模

在实际生活中，不同用户存在不同的生活习惯，有的用户喜欢在白天出去并签到，有的用户喜欢在晚上出去并签到，这些签到行为往往反映出他们的个人喜好以及生活习惯。因此，两个用户签到行为分布越相似，则他们越可能具有相同的个人爱好以及行为习惯，根据同质性理论[15]，则可以认为他们成为好友链接的可能性越大。

首先，用户在一天 24 小时均存在出去签到的可能，为了使概率函数不存在严重偏差，本文将用户签到时间分为 24 等分，对应一天的 24 个时间槽，然后统计用户在这 24 个时间槽的签到频率。其中，本文统计了用户 u 和用户 v 的签到时间频率，结果如图 1 所示。其中，直方图分别表示了用户 u 和用户 v 的签到时间频率分布情况。从图中可以看出用户 u 和 v 的签到习惯有着较为明显差异，因此，可以认为二者在签到行为习惯上存在较低的相似性。



(a)用户 u 的签到分布



(b)用户 v 的签到分布

图 1　用户签到分布

Fig. 1　User sign-in distribution

上述方法虽然可以判断用户之间的相似性，但是却难以达到理想的结果，因为将一天划分为 24 个时间槽，这样会导致时间槽内不同的时间点，签到频率却相同，这显然不符合逻辑。因此，可以采用高斯核函数建立用户基于签到时间的核密度估计分布，如图 1 中的曲线。使用连续分布表示优势在于:a)可以准确的反映出用户在一天中连续时间下的签到行为分布；b)对于签到比较稀疏的用户，他们在大部分的时间槽中均不存在签到行为，如若采用离散统计的思想计算用户之间的相似性，会造成严重的偏差，因而通过连续的核密度估计分布可以有效缓解该问题。

假设用户 u 在时间段 s 的签到概率为 $P_{t,u}(s)$ ，在一维核密度估计下，通过余弦相似度函数得到两个用户的相似值 $\mathrm{sim}_t(u,v)$ ，公式如下：

$$\mathrm{sim}_t(u,v)=\cos(u,v)=\frac{\sum_{s\in S_{uv}}P_{t,u}(s)P_{t,v}(s)}{\sqrt{\sum_{s\in S_{uv}}P_{t,u}^2(s)\sum_{s\in S_{uv}}P_{t,v}^2(s)}} \tag{8}$$

其中: S_{uv} 为签到时间集合。

2.3 用户个性化空间建模

本文采用二维核密度估计方式，以挖掘单个用户对新的位置签到的概率。令 $S_u=(p_1,p_2,\ldots,p_n)$ ，表示用户 u 访问的位置集合。利用二维核密度估计方式获得用户 u 访问某一个新位置 p 的概率 $P(p\mid S_u)$ ：

$$P(p\mid S_u)=\frac{1}{n\sigma^2}\sum_{i=1}^{n}\mathrm{K}(\frac{p-p_i}{\sigma}) \tag{9}$$

其中: $p_i=(\mathrm{lat}_i,\mathrm{lon}_i)^{\mathrm{T}}$ 表示位置 p_i 的二维空间向量坐标， lat_i 表示经度， lon_i 表示纬度。 $\mathrm{K}(\cdot)$ 表示核函数， σ 表示平滑窗口，也称为窗宽。

式(9)中，核函数选择的是标准的高斯核函数，表示如下：

$$\mathrm{K}(x)=\frac{1}{2\pi}\exp(-\frac{1}{2}x^{\mathrm{T}}x) \tag{10}$$

最优窗宽设定为 $\sigma=\mathrm{n}^{-\frac{1}{6}}\sqrt{\frac{1}{2}\hat{\sigma}^{\mathrm{T}}\hat{\sigma}}$ ， $\hat{\mu}$ 和 $\hat{\sigma}$ 分别表示 S_u 集合中经度值和纬度值的均值和方差，计算公式如下：

$$\hat{\sigma}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-\hat{m})^2} \tag{11}$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}p_i \tag{12}$$

3 协作式个性化链接预测算法

3.1 社团划分

传统社交网络中社团划分采用基于网络结构的划分方式，本文为了更好地挖掘相同兴趣群体的用户，利用上述兴趣组的定义为用户划分社团。如图 2 所示，由于一个用户可能会访问多个位置，而且同一个位置也可能属于多个类别。因此，基于兴趣组划分出来的社团为一个重叠社团，即一个用户可能属于多个社团。基于兴趣组的划分方式，将 G_h 划分为了 $K=|\mathrm{C}|$ 个社团，则可以构造一个 $N*K$ 的用户-社团矩阵，记作 F 。 $F_{uc}=1$ 表示用户 u 属于社团 c ，反之， $F_{uc}=0$ 表示用户 u 不属于社团 c 。

3.2 用户链接预测

如果用户 u 和用户 v 属于同一个社团，则表明他们二者存在相似的兴趣爱好，当他们拥有的共同社团越多，则相似度也就越大，从而产生链接的可能性也越大。本文根据文献[16]中的方法计算社团 c 中的两个用户 u,v 的链接概率为

$$P_{uv}^{\mathrm{U}}(c)=1-\exp(-F_{uc}\cdot F_{vc}) \tag{13}$$

如果用户 u,v 中有一个不属于社团 c ，则 $F_{uc}=0$ 或 $F_{vc}=0$ ，且 $P_{uv}^{\mathrm{U}}(c)=0$ 。由于用户可能会属于多个社团，则 u,v 不存在链接的概率可表示为

$$1-P_{uv}^{\mathrm{U}}=\prod_{c}(1-P_{uv}^{\mathrm{U}}(c))=\exp(-\sum_{c}F_{uc}\cdot F_{vc}) \tag{14}$$

其中： u,v 存在链接的概率 $P_{uv}^{\mathrm{U}}(c)$ 为

$$P_{uv}^{\mathrm{U}}(c)=1-\exp(-\sum_{c}F_{uc}\cdot F_{vc}) \tag{15}$$

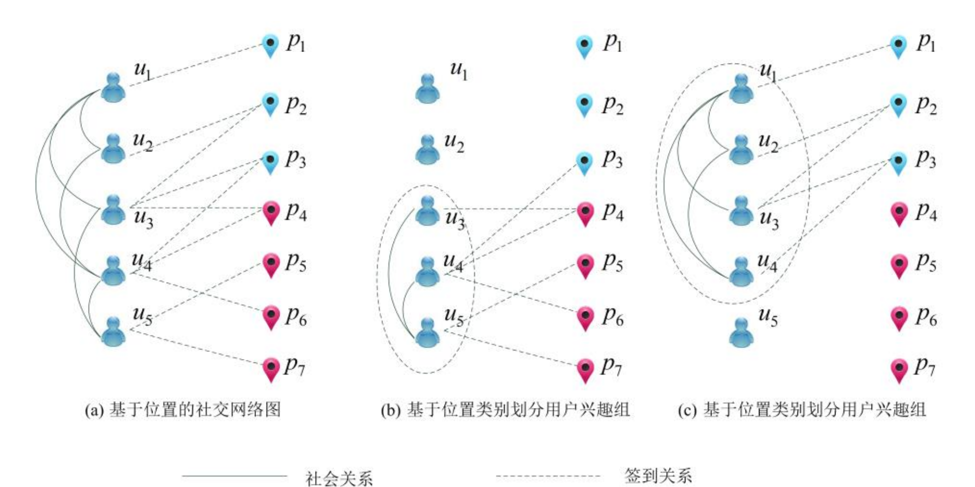


图 2　基于兴趣组划分社团

Fig. 2　Based on interest groups to divide community

由于用户共同签到的位置个数以及用户共同好友的个数也对链接存在概率有影响。因此，可以将式(15)做如下修正：

$$P_{\mathrm{uv}}^{\mathrm{U}}(c)=1-\partial\cdot\exp(-\sum_{u'}U_{\mathrm{uu'}}^{\mathrm{V}}\cdot U_{\mathrm{vu'}}^{\mathrm{V}})-\beta\cdot\exp(-\sum_{c}F_{\mathrm{uc}}\cdot F_{\mathrm{vc}})-(1-\partial-\beta)\cdot\exp(-\sum_{p}P_{\mathrm{up}}^{\mathrm{V}}\cdot P_{\mathrm{vp}}^{\mathrm{V}})\tag{16}$$

其中：$\boldsymbol{U}^{\mathrm{V}}$ 表示用户-用户关系矩阵，$U_{\mathrm{uu'}}^{\mathrm{V}}=1$ 表示用户 u 和 u' 存在好友关系；$\boldsymbol{P}^{\mathrm{V}}$ 表示用户-位置关系矩阵，$P_{\mathrm{up}}^{\mathrm{V}}=1$ 时表示用户 u 在位置 p 有过签到，通过以上方式可以得到网络中每个用户与其他用户产生链接的概率。在此基础上，考虑用户之间关于签到时间分布的相似性 $\mathrm{sim}_t(u,v)$，将用户个性化的行为习惯进行匹配，对每个用户与其他非好友用户的链接存在概率进行更新，更新方式如下：

$$\tilde{P}_{\mathrm{uv}}^{\mathrm{U}}=P_{\mathrm{uv}}^{\mathrm{U}}\cdot\mathrm{sim}_t(u,v)\tag{17}$$

其中：$\mathrm{sim}_t(u,v)$ 表示用户基于签到时间的相似度。

3.3　位置链接预测

为合理预测用户与位置之间的链接关系，本文首先作出以下假设：给定一个目标用户 $u\in c_k$，则认为用户 u 会更愿意访问对于社团 c_k 来说重要的位置，如果用户 u 属于多个社团，则用户更愿意访问的位置由多个社团综合决定。

为了找到社团 c_k 中的重要位置，本文采用从社团 c_k 重启的随机游走，如下所示：

$$r_{c_k,u}^{(t)}=(1-\zeta)\sum_{u'\in N_u(u)}w(u',u)r_{c_k,u'}^{(t-1)}+(1-\zeta)\sum_{p'\in N_p(u)}w(p',u)r_{c_k,p'}^{(t-1)}+\zeta\cdot\frac{1}{|c_k|}r_u^{c_k}\tag{18}$$

$$r_{c_k,p}^{(t)}=(1-\zeta)\sum_{u'\in N_u(p)}w(u',p)r_{c_k,u'}^{(t-1)}\tag{19}$$

其中：$r_{c_k,u}^{(t)}$ 表示用户节点的到达概率，$r_{c_k,p}^{(t)}$ 表示位置节点的到达概率，t 表示迭代次数，$N_u(u)$ 为用户 u 的邻居用户集合，$N_p(u)$ 为用户 u 的邻居位置集合，$N_u(p)$ 为位置 p 的邻居用户集合。$w(u',u)$ 为用户节点之间的转移概率，$w(p',u)$ 为位置 p' 到用户 u 的转移概率，$w(u',p)$ 为用户 u' 到位置 p 的转移概率。ζ 为随机游走的重启概率，如果 u 属于社团 c_k，$r_u^{c_k}=1$，反之 $r_u^{c_k}=0$。

因为位置节点仅与用户相连，所以位置节点与用户节点之间的转移概率可以表示为

$$w(p,u)=\frac{1}{N_u(p)}\tag{20}$$

用户节点可与位置节点或用户节点相连，为了协调用户节点与位置节点和用户节点的权重关系，引入调节参数 λ。由于存在两种类型的边，且本文认为同种类型边之间，其转移概率相同，则两个用户节点之间的转移概率可以设置如下：

$$w(u',u)=\begin{cases}\dfrac{(1-\lambda)P_{\mathrm{u'u}}}{\sum_{i\in N_u(u')}P_{\mathrm{u'i}}}, & 若|N_u(u')|>0且|N_p(u')|>0\\ \dfrac{P_{\mathrm{u'u}}}{\sum_{i\in N_u(u')}P_{\mathrm{u'i}}}, & 若|N_u(u')|>0且|N_p(u')|=0\\ 0 & 其他\end{cases}\tag{21}$$

当 $(u',u)\in E$ 时，$P_{\mathrm{u'u}}=1$，否则 $P_{\mathrm{u'u}}$ 等于 (u',u) 的存在概率 $\tilde{P}_{\mathrm{u'u}}^{\mathrm{U}}$。

由于不同位置对于用户的重要程度不同，本文需要量化位置 p 对于用户 u 的重要性 $w_u(p)$，本文主要采用定义的本地位置重要度和全局位置重要度来量化该指标，即

$$w_u(p)=\mathrm{LLI}_u(p)\cdot\mathrm{GLI}_u(p)\tag{22}$$

则用户到位置的转移概率可以表示如下：

$$w(u,p)=\begin{cases}\dfrac{\lambda w_u(p)}{\sum_{p'\in N_p(u)}w_u(p')}, & 若|N_p(u)|>0且|N_u(u)|>0\\ \dfrac{w_u(p)}{\sum_{p'\in N_p(u)}w_u(p')}, & 若|N_p(u)|>0且|N_u(u)|=0\\ 0 & 其他\end{cases}\tag{23}$$

当基于社团 c_k 重启的随机游走达到收敛时，该社团下各位置节点的到达概率 $\boldsymbol{r}_{c_k,p}$ 可以理解为在社团 c_k 下各位置的重要程度，如下所示：

$$\boldsymbol{r}_{c_k,p}=[r_{c_k,p_1},r_{c_k,p_2},\cdots,r_{c_k,p_n}]\tag{24}$$

由于网络中存在 K 个社团，所以可以得到社团-位置矩阵 $\boldsymbol{A}^{\mathrm{C}}$，如下所示：

$$\boldsymbol{A}^{\mathrm{C}}=[\boldsymbol{r}_{c_1,p}\ \boldsymbol{r}_{c_2,p}\ \cdots\ \boldsymbol{r}_{c_k,p}]^{\mathrm{T}}\tag{25}$$

利用社团-位置矩阵 $\boldsymbol{A}^{\mathrm{C}}$ 和用户-社团关系 $\boldsymbol{F}$ 的乘积表示在社团驱动下的用户 u 访问位置 p' 的概率，计算公式如下：

$$P_{\mathrm{up'}}^{\mathrm{V}}=\sum_{c_i\in C}F_{\mathrm{uc}_i}\cdot A_{\mathrm{c}_i\mathrm{p'}}^{\mathrm{C}}\tag{26}$$

其中：C 表示网络中的所有社团，$A_{\mathrm{c}_i\mathrm{p'}}^{\mathrm{C}}$ 表示位置 p' 在社团 c_i 中的重要程度。

上述通过社团关系选出的重要位置，是从与目标用户具有相似兴趣的样本用户集合中总结出的位置，是从兴趣层面得到的。然而，由于地理位置的影响，用户并不一定会去访问这些位置，例如，用户 u 是一个宅男，很少访问较远的位置，而用户 v 喜欢旅游，经常环游各国，如果从兴趣层面发现用户 u 和 v 都对位置 p 比较感兴趣，且位置 p 离用户 u，v 的距离均较远，考虑到用户 u 可能对于距离的容忍度低于用户

v，则可以认为用户 u 访问位置 p 的概率会小于用户 v。因此，预测一个用户是否会去访问某个位置需要同时结合兴趣层面和每个用户的位置空间容忍度考虑。

本文通过用户的个性化空间建模，得到用户在自身空间容忍度下访问位置 p 的概率 $P(p|S_u)$，从而更新用户 u 从兴趣层面得到的访问概率矩阵，更新公式如下：

$$\tilde{P}_{up}^{V}=P_{up}^{V}\cdot P(p|S_u) \quad (27)$$

其中：$\tilde{P}_{up}^{V}$ 表示用户 u 可能会访问 p 的概率，依据 $\tilde{P}_{up}^{V}$ 更新 $\boldsymbol{P}^{V}$ 矩阵，如下：

$$(P_{up}^{V})^{t}=\begin{cases}\tilde{P}_{up}^{V}, & 若(P_{up}^{V})^{t-1}=0\\ (P_{up}^{V})^{t-1}, & 其他\end{cases} \quad (28)$$

其中：$(P_{up}^{V})^{t}$ 是迭代 t 次后用户 u 访问位置 p 的概率。对 $\boldsymbol{P}^{V}$ 矩阵更新之后，就可以开始新一轮的用户链接预测以及位置链接预测过程。本文的算法描述如算法 1 所示。

文中用户链接预测算法复杂度为 $O((a+b+m)n^2)$，其中 a 为共邻用户数，b 为社团数，m 为共邻位置数，n 为用户数。位置链接预测算法复杂度为 $O(bmt_1(m+n))$，t_1 为随机游走收敛次数。假设相互迭代次数为 t_2，则 CPP 算法的时间复杂度为 $O(t_2((a+b+m)n^2+bmt_1(m+n)))$。

算法 1：CPP 算法

输入：基于位置的社交网络 $G_h=(U_h,P_h,\boldsymbol{E}_h,W_h)$；初始预测空间 C。

输出：网络中可能存在的用户链接 E_U^{+} 以及位置链接 E_P^{+}。

1. 基于单个用户的签到时间，采用一维核密度估计方式建模用户的签到行为概率分布 P_t，进而得用户之间的相似度 $\mathrm{sim}_t(u,v)$；
2. 基于用户历史访问位置的经纬度信息，采用二维核密度估计方式建模用户的空间容忍度信息，得到每个用户访问新位置的概率 $P(p|S_u)$；
3. 基于兴趣组定义划分用户重叠社团 C，构建用户-社团矩阵 $\boldsymbol{F}$；
4. repeat
5. //用户链接预测
6. 依据式(13)计算用户之间链接的存在概率 P_{uv}^{U}；
7. 依据式(14)更新用户链接概率 P_{uv}^{U}，得到个性化用户链接概率 $\tilde{P}_{uv}^{U}$；
8. //位置链接预测
9. 依据式(17)~(20)计算网络中的边权值;
10. 依据式(15) (16)计算每个社团中各位置的到达概率 $r_{c_i,p}$；
11. 依据式(23)计算用户与位置之间链接存在的概率 P_{up}^{V}；
12. 依据式(24)更新位置链接概率 P_{up}^{V}，得到个性化的位置链接概率 $\tilde{P}_{up}^{V}$；
13. 根据式(25)更新用户-位置矩阵 $\boldsymbol{P}^{V}$；
14. until 达到指定迭代次数
15. 通过 $\tilde{\boldsymbol{P}}^{U}$ 和 $\boldsymbol{P}^{V}$ 矩阵得到网络中可能存在用户链接 E_U^{+} 和位置链接 E_P^{+}；

4 仿真结果与分析

4.1 数据集描述

本文采用从 Gowalla 抓取到的数据集进行实验，数据集来源于文献[17]，包含了用户表、位置表、好友关联表以及签到表。其中用户表包含了用户签到的次数以及签到的位置类别数等信息；位置表中包含了位置的经纬度信息、所属城市以及所属类别；好友关系表包含了用户之间的好友连边关系；签到表包含了用户签到的位置以及对应的签到时间，时间精确到小时。

实验中，本文从原始数据中选取在柏林和休斯顿的签到记录作为实验数据集，删除总签到次数小于 10，或签到位置数不大于 2 的不活跃用户，以及删除被签到次数小于 5，或被签到用户数不大于 2 的位置，减小数据稀疏对实验的影响。表 1 为处理后的数据集详细描述。

表 1 实验数据统计

Table 1 Experimental data statistics

数据集	用户数	位置数	签到数	关系数	类别数
柏林	5510	15528	238972	29207	71
休斯顿	11138	29383	512977	61221	83

4.2 实验结果分析

在实验中，首先设定用户链接预测以及位置链接预测最大迭代次数为 30 次，重启随机游走的重启概率 $\varsigma=0.8$。调节参数 λ 变化对比如图 3 所示，不同调节参数影响位置链接预测准确率，在 $\lambda=0.4$ 左右时，位置链接预测准确率最优。用户链接预测参数 α,β 用于调节用户的共同社团，用户共同好友个数以及用户共同签到位置个数的权重对用户链接预测的影响。如表 2 所示，在实际数据集中验证参数 α,β 对用户链接预测的影响，在 β 相对增大时，链接预测准确率提高，且在 $\alpha=0.3$，$\beta=0.4$ 左右时，用户链接预测准确率最优。因此，设置柏林数据集中用户链接预测参数 $\alpha=0.3$，$\beta=0.4$，休斯顿数据集 $\alpha=0.3$，$\beta=0.5$。仿真主要考虑了两个因素对实验结果的影响，一是不同比例训练样本对不同方法的影响；二是迭代次数对本文提出 CPP 算法的影响。

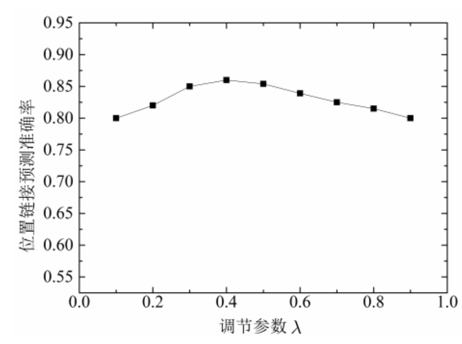


图 3 不同调节参数位置链接预测准确率

Fig. 3 Positional link prediction accuracy of different parameters

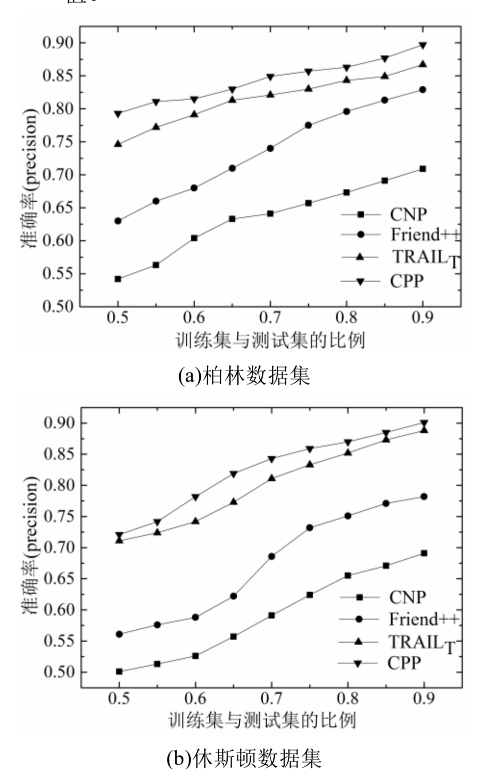
表 2 不同 α 和 β 用户链接预测准确率

Table 2 User link prediction accuracy of different α and β

β	α = 0.1	0.2	0.3	0.4	0.5
0.1	0.536	0.556	0.589	0.576	0.569
0.2	0.589	0.676	0.728	0.702	0.698
0.3	0.661	0.716	0.778	0.755	0.731
0.4	0.695	0.767	0.851	0.835	0.824
0.5	0.686	0.758	0.849	0.828	0.807

本文选取了典型链接预测方法与 CPP 算法进行比较，对比算法包括：位置的共同邻居(Common Neighbors of Places, CNP)[18]是一种用户链接预测算法，如果两个用户有更多的共同朋友访问被其中一个用户访问过的地方，则他们之间产生链接的概率越大；改进的带重启的随机游走算法 Friend++[19]也是一种用户链接预测算法，该算法是改进了传统重启随机游走算法，将加权平均方法集成到随机游走方法中，预测用户链接概率；Rank-GeoFM[20]，位置链接预测算法，通过用户偏好、签到位置以及时间地理上下文的影响优化位置排序函数，得到最优位置链接；TRAIL$_T$[11]，同时预测用户链接以及位置链接的算法，通过最大化用户链接概率函数和位置链接概率函数的乘积，来同时获取最优用户链接以及最优位置链接。

首先，比较不同比例训练样本对算法性能的影响，本文主要采用 $<0.5, 0.55, \cdots, 0.9>$ 这 9 种标准来划分数据集。为了保证实验的可靠性，本文取 10 次实验结果的平均值作为最终实验结果，其中用户链接预测的结果如图 4、5 所示。图 4 表示用户链接预测准确率的实验结果，其中横坐标表示训练集占比情况，纵坐标表示算法准确率。图 5 表示用户链接预测 AUC 值的实验结果，其中横坐标表示训练集占比情况，纵坐标表示算法 AUC 值。

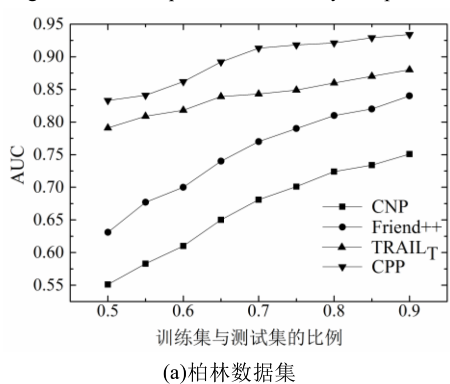


(a)柏林数据集

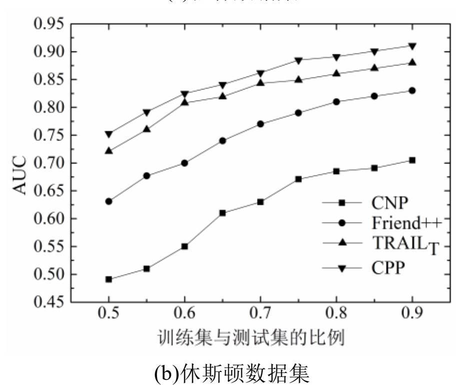
(b)休斯顿数据集

图 4 用户链接预测准确率对比

Fig. 4 User link prediction accuracy comparison



(a)柏林数据集



(b)休斯顿数据集

图 5 用户链接预测 AUC 值对比

Fig. 5 User prediction AUC comparison

AUC 直接物理意义是 ROC 曲线下的面积，ROC 曲线是分类器常用的性能评价指标。因为一个二分类器在输出结果是 1 还是 0，往往会取决于输出概率和预设的概率阈值，概率阈值的选取一定程度影响分类器的性能，为了无论阈值怎么选取，分类器评价指标尽可能的正确，采用了 ROC 曲线这种衡量指标，但是 ROC 曲线只是反映了分类器的分类能力，通过 AUC 值可以量化 ROC 曲线，直观呈现分类器的能力，AUC 值越大分类效果越好，AUC 的理想值为 1。

从图 4、5 可以看出，随着训练集比例的增大，所有方法的准确率以及 AUC 值都处于上升趋势。且 CPP 算法的性能始终优于 TRAIL 算法，原因在于 TRAIL 算法未能有效把握用户个性化特征，忽略了对用户之间行为习惯相似性和用户空间位置容忍度的考虑，而 CPP 在算法迭代的基础上充分融合了用户在时间和空间上的个性化特征，使得预测结果得以提升。CPP 算法和 TRAIL 算法性能明显优于 CNP 和 Friend++，原因在于 Friend++算法仅利用到了用户的社交关系，而没有考虑用户的签到关系，所以对于信息的利用不够完善，同时 CNP 以及 Friend++算法未能融入迭代思想，从而导致算法性能不太理想。

表 3 位置链接预测各项性能指标对比

Table 3 Location link prediction performance indicators comparison

数据集		柏林数据集训练集比例					休斯顿数据集训练集比例				
指标	方法	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
AUC	Rank-GeoFM	0.652	0.671	0.703	0.721	0.755	0.635	0.646	0.688	0.717	0.739
	$TRAIL_T$	0.613	0.683	0.719	0.748	0.765	0.623	0.643	0.679	0.714	0.746
	CPP	0.715	0.744	0.785	0.825	0.842	0.706	0.733	0.766	0.795	0.814
准确率	Rank-GeoFM	0.611	0.632	0.667	0.684	0.705	0.591	0.617	0.634	0.659	0.689
	$TRAIL_T$	0.597	0.628	0.663	0.698	0.723	0.598	0.623	0.645	0.653	0.708
	CPP	0.709	0.734	0.779	0.816	0.857	0.699	0.716	0.742	0.779	0.807

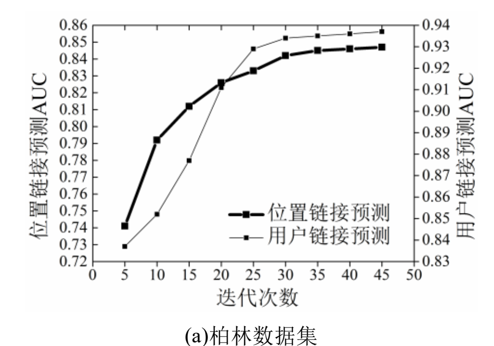
位置链接预测的实验结果如表 3 所示。从表中可以看出，休斯顿数据集中所有算法的准确率以及 AUC 值都偏低，原因是休斯顿数据集偏大，网络中的数据稀疏性较强，导致算法预测性能下降。同时可以看出，CPP 算法的预测性能始终最优，因为 CPP 算法集合了社团知识以及个性化选择，所以可以合理挖掘出用户的兴趣爱好，准确预测网络中可能的位置链接。

最后，为了验证算法的迭代可以有效提高链接预测的性能，设置训练集比例为 0.9，迭代次数分别设置为 $<5, 15, \cdots, 45>$ 对每个迭代次数重复 10 次实验取平均值作为预测结果。实验结果如图 6 所示，可以明显看出，随着迭代次数的增加，用户链接预测以及位置链接预测的 AUC 值均不断提升，当迭代次数到达 30 次左右时，AUC 值趋于平稳。实验证明了 CPP 算法的迭代过程能够有效提升链接预测的性能。

5 结束语

本文解决了传统的链接预测方法独立求解用户链接以及位置链接，同时未能把握用户在时间和空间上的个性化等问题。提出了一种 LBSN 中协作式个性化链接预测(CPP)算法，通过一维和二维核密度估计方式分别对每个用户的签到时间以及签到空间进行建模，挖掘用户的签到行为习惯以及空间位置容忍度。同时基于重叠社团重启的随机游走方式，将用户的个性化特征有效的融入到模型中，并通过有限的迭代过程使得两种预测性能不断提升，同时达到最优。本文方法从一种新的角度同时完成用户链接预测以及位置链接预测任务，有效挖掘和利用了两种任务之间的相关性，不足之处在于算法的时间复杂度较大，下一步的计划是通过算法优化解决

CPP 中时间复杂度高的问题。



(a)柏林数据集

(b)休斯顿数据集

图 6　AUC 值随迭代次数变化对比

Fig. 6　AUC comparison with the number of iterations

参考文献：

[1] Stepan T, Morawski J M, Dick S, *et al.* Incorporating spatial, temporal, and social context in recommendations for location-based social networks [J]. IEEE Trans on Computational Social Systems, 2016, 3 (4): 164-175.

[2] 文馨, 陈能成, 肖长江. 基于 Spark GraphX 和社交网络大数据的用户影响力分析 [J]. 计算机应用研究, 2018, 35 (3): 830-834. (Wen Xin, Chen Nengcheng, Xiao Changjiang. Analysis of user influence based on social network big data and Spark GraphX [J]. Application Research of Computers, 2018, 35 (03): 830-834.)

[3] Zhou Ningnan, Zhao Waynexin, Zhang Xiao, *et al.* A general multi-context embedding model for mining human trajectory data [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28 (8): 1945-1958.

[4] Xiao Yunpeng, Li Xixi, Wang Haoxun, *et al.* 3-HBP: a three-level hidden bayesian link prediction model in social networks [J]. IEEE Trans on Computational Social Systems, 2018, 5 (2): 430-443.

[5] Valverda-rebaza J, Roche M, Poncelet P, *et al.* The role of location and social strength for friendship prediction in location-based social networks [J]. Information Processing & Management, 2018, 54 (4): 475-489.

[6] 丁勇, 刘菁, 蒋翠清, 等. LBSN 中考虑用户交友偏好的好友推荐方法研究 [J]. 系统工程理论与实践, 2017, 37 (11): 2975-2982. (Ding Yong, Liu Jing, Jiang Cuiqing, *et al.* A study of friends recommendation algorithm considering users’ preference of making friends in the LBSN [J]. Systems Engineering Theory & Practice, 2017, 37 (11): 2975-2982.)

[7] Bayrak A E, Polat F. Examining place categories for link prediction in location based social networks [C]// Proc of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. NewYork: ACM Press, 2016: 976-979.

[8] Pavlos K, Yannis M. A time-aware Spatio-textual Recommender system [J]. Expert Systems with Applications, 2017, 78: 396-406.

[9] 李鑫, 刘贵全, 李琳, 等. LBSN 上基于兴趣圈中社会关系挖掘的推荐算法 [J]. 计算机研究与发展, 2017, 54 (2): 394-404. (Li Xin, Liu Guiquan, Li Lin, *et al.* Circle-based and social connection embedded recommendation in LBSN [J]. Journal of Computer Research and Development, 2017, 54 (2): 394-404.)

[10] Hosseini S, Li L T. Point-of-interest recommendation using temporal orientations of users and locations [C]//Proc of International Conference on Database Systems for Advanced Applications. Dallas: Springer International Publishing, 2016: 330-347.

[11] Zhang Jiawei, Kong Xiangnan, Yu P S. Transferring Heterogeneous Links Across Location-based Social Networks [C]// Proc of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2014: 303-312.

[12] Zhan Qianyi, Zhang Jiawei, Yu P S. Integrated anchor and social link predictions across multiple social networks [J]. Knowledge and Information Systems, 2018, 6: 1-24.

[13] Wang Hao, Terrovitis M, Mamoulis N. Location recommendation in Location-based Social Networks using user Check-in Data [C]// Proc of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2013: 374-383.

[14] Yao Lina, Sheng Quan, Wang Xianzhi, *et al.* Collaborative location recommendation by integrating multi-dimensional contextual information [J]. ACM Trans on Internet Technology, 2018, 18 (3): 1-24.

[15] Timothy L F, Neville J. Randomization tests for distinguishing social influence and homophily effects [C]// Proc of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 601-610.

[16] Yang J, Mcauley J, Leskovec J. Community Detection in Networks with Node Attributes [C]//Proc of the 13th IEEE International Conference on Data Mining. Piscataway,NJ: IEEE Press, 2013: 1151-1156.

[17] Liu Yong, Wei Wei, Sun Aixin, *et al.* Exploiting geographical neighborhood characteristics for location recommendation [C]// Proc of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2014: 739-748.

[18] Alverda-rebaza J, Roche M, Poncelet P, *et al.* Exploiting social and mobility patterns for friendship prediction in location-based social networks [C]// Proc of the 23rd International Conference on Pattern Recognition. Piscataway,Nj: IEEE, 2016: 2526-2531.

[19] Gong Jibing, Gao Xiaoxia, Cheng Hong, *et al.* Integrating a weighted-average method into the random walk framework to generate individual friend recommendations [J]. Science China Information Sciences, 2017, 60 (11): 110104.

[20] Li Xutao, Gao Cong, Li Xiaoli, *et al.* Rank-geofm: a ranking based geographical factorization method for point of interest recommendation [C]// Proc of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2015: 433-442.